Journal of Quantitative Criminology, Vol. 16, No. 3, 2000

Developing a Multicenter Randomized Trial in Criminology: The Case of HIDTA

David Weisburd[1] and Faye S. Taxman[2]

In criminal justice, as in other fields, an experimental study conducted at a single site does not offer a solid basis upon which to make strong public policy recommendations. To date, criminal justice researchers have relied upon two general approaches to overcome the limitations of single-site experimental research. The first, termed "meta-analysis," seeks to combine independent studies to identify consistent effects across criminal justice settings or contexts. The second, sometimes termed "replication studies," seeks to replicate investigations in multiple criminal justice jurisdictions. In this paper we describe a related approach developed in clinical studies in medicine and examine its applicability in criminal justice settings. Termed a "multicenter clinical trial," this method demands the implementation of a single experimental protocol at multiple sites. We contrast the multicenter approach with other methods and provide a substantive example of an ongoing multicenter criminal justice study. We begin by examining the specific limitations of current approaches and solutions offered by multicenter studies to overcome these. We then turn to an application of the multicenter clinical trial in a criminal justice setting. Using the example of the HIDTA (High-Intensity Drug Trafficking Areas) evaluation of drug treatment programs currently being conducted at multiple sites, we illustrate components of the multicenter approach as well as potential drawbacks researchers are likely to face in its application in crime and justice studies.

KEY WORDS: randomized experiments; multicenter randomized trial; drug treatment; criminal justice; meta-analysis; replication studies; statistical power.

1. INTRODUCTION

The limitations of single-site evaluation studies are well recognized by criminal justice evaluators. In part, it is the significant variation that is found across criminal justice agencies that suggests caution in developing broad public policy recommendations from single-site studies. A finding of

[1]Department of Criminology and Criminal Justice, University of Maryland, College Park, Maryland 20742; Institute of Criminology, The Hebrew University, Jerusalem, Israel.
[2]Department of Criminology and Criminal Justice, University of Maryland, College Park, Maryland 20742.

0748-4518/00/0900-0315$18.00/0

no impact at one site may have more to do with the particular characteristics of the agency involved than the strengths of the program proposed. Similarly, a strong impact of treatment in one jurisdiction may not carry over to others that have offenders drawn from different ethnic communities or that come from different socioeconomic backgrounds (e.g., see Berk *et al.*, 1992b; Sherman *et al.*, 1992). In criminal justice, as in other fields, a study conducted at a single site does not offer a solid basis upon which to make strong public policy recommendations (Cook *et al.*, 1992; Garner, 1990; Greenberg *et al.*, 1994; MacKenzie and Souryal, 1994; McShane *et al.*, 1992; Tilly, 1994).

The difficulty of coming to solid conclusions about criminal justice policies from single-site evaluations has been of particular concern in experimental studies. Randomized experiments allow researchers to make a stronger link between interventions and program outcomes than nonexperimental studies (Campbell and Stanley, 1966; Sechrest and Rosenblatt, 1987). As a result, randomized experiments are often given greater weight than nonexperimental studies in developing public policy (e.g., see Farrington *et al.*, 1986). But the strength of experimental designs in specifying treatment impacts for specific populations does not in itself overcome the weaknesses associated with single site research studies (e.g., see Binder and Meeker, 1988). The case of the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984) provides a well-known example of this problem. While the presence of a deterrent effect for arrest in Minneapolis helped to reinforce the development of mandatory arrest policies in police agencies throughout the country (Sherman and Cohen, 1989; U.S. Attorney General, 1984), subsequent replications of the study in other cities found arrest policies to be ineffective and sometimes harmful (Garner *et al.*, 1995; Hutchison and Hirschel, 1994).

To date criminal justice researchers have relied upon two general approaches to overcome the limitations of single-site experimental research. The first, termed "meta-analysis" (see Glass, 1976; Glass *et al.*, 1981; Cook *et al.*, 1992), seeks to combine independent studies to identify consistent effects across criminal justice settings or contexts. The second, which Garner (1990) defines as "replication studies," seeks to replicate investigations in multiple criminal justice jurisdictions. In this paper we describe a related approach developed in clinical studies in medicine and examine its applicability in criminal justice settings. Termed a "multicenter clinical trial," this method demands the implementation of a single experimental protocol at multiple sites (see Borok *et al.*, 1994; Fleiss, 1982; Friedman *et al.*, 1985; Hill, 1962; Stanley *et al.*, 1981). The multicenter trial enables the researcher to develop a statistically powerful research design for analyzing overall treatment impacts at the same time that it provides a method for taking into account the importance of intersite variation on experimental outcomes.

Below we contrast the multicenter approach with other methods and provide a substantive example of an ongoing multicenter criminal justice study. We begin by examining the specific limitations of current approaches and the ways in which multicenter studies seek to overcome these. We then turn to an application of the multicenter clinical trial in a criminal justice setting. Using the example of the Washington/Baltimore HIDTA Project and Multi-Center Evaluation of drug treatment programs, we illustrate components of the multicenter approach as well as potential drawbacks researchers are likely to face in its application in crime and justice settings.

2. LIMITATIONS OF CURRENT METHODS

Meta-analysis is a common method used for overcoming the limitations of single-site evaluations (Cook *et al.*, 1992; Glass *et al.*, 1981; Light and Pillemer, 1984). Glass (1976) coined the term "meta-analysis" to describe the pooling of multiple studies in a specific area of interest into a single analysis in which each study is an independent observation. Accordingly, in a meta-analysis the researcher selects all program evaluations that fit a set of a priori criteria and then analyzes the results using each study outcome as a single data point in a larger analysis of program effects (for examples in criminal justice, see Andrews *et al.*, 1990; Petrosino, 1995; Whitehead and Lab, 1989). The main advantage of meta-analysis is that it allows the development of a single estimate of effect size over a large number of studies. Because this estimate is an average of estimates over a series of investigations, the researcher avoids the fallacy of placing too much weight on a single study.

While meta-analysis can provide a straightforward solution to the problem of drawing inferences from single-site experimental studies, it is often criticized for its inclusion of a diverse set of investigations within a single analytic context (e.g., see Logan and Gaes, 1993; Eysenck, 1978). In principle, a meta-analysis may select studies using very specific and restricted criteria. And indeed, researchers who use meta-analytic techniques recommend the development of clear and precise rules for the inclusion of studies (e.g., see Abrami *et al.*, 1988; Petrosino, 1995; Sherman *et al.*, 1997). However, in practice meta-analyses often include studies not only from different jurisdictions, but that rely upon methods of sampling, measurement, and data collection that vary greatly in style and quality.

This problem is particularly acute in meta-analysis of experimental research. Unlike fields such as medicine, in which there are thousands of randomized experiments conducted each year and more than 250,000 experiments reported (Sakett and Rosenberg, 1995), experimental criminal justice studies are relatively rare (see Sherman, 1998). Even the most optimistic reviews of randomized studies in criminal justice number the total

number of criminal justice experiments in the hundreds (e.g., see Petrosino, 1997). Researchers, accordingly, must take a fairly broad approach if they want to gain enough studies to come to solid conclusions about program impacts. For example, in his study of juvenile delinquency treatment, Lipsey (1992, pp. 87–88) includes studies that have "some intervention or treatment, *broadly defined*, that had as its aim (explicitly or implicitly) the reduction, prevention, treatment remediation, and so forth, of delinquency or antisocial behavior problems similar to delinquency." Meta-analyses are not generally used to define the advantages of specific programs or approaches but, rather, to define general effects within a broad area of inquiry (Petrosino, 1995).[3]

Even if a sufficient number of experimental studies regarding a specific type of treatment could be identified for meta-analysis, it is likely that there would be significant variation in the methods used by researchers. Meta-analysis does not require that investigations are coordinated in terms of research protocols. Indeed, meta-analysis was developed primarily as a method of grouping independent studies into a single analytic framework. But without this coordination, it is very likely that there will be substantive differences in the nature of the treatments implemented and basic components of the study design.

"Replication" studies provide an approach for overcoming the weaknesses of single-site experimental evaluations that allows for greater consistency in method and research design (Garner, 1990). In this case, studies are replicated at multiple sites either by single research teams or by multiple research teams working in coordination with each other (e.g., see Petersilia and Turner, 1993a, b; Schneider, 1986; Sherman, 1992). Researchers begin with a common program or public policy approach which is drawn either directly or by inference from prior studies. For example, in the Spouse Assault Replication Program [Garner *et al.*, 1995; National Institute of Justice (NIJ), 1987], the National Institute of Justice supported six replications conducted by six independent research teams. These replications were based directly upon one prior study, the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984). In contrast, in the RAND intensive probation studies, one group of investigators, Petersilia and Turner (1993a, b), sought to test the effects of intensive probation in separate experimental evaluations across 11 states. This study, in turn, was based on a general public policy approach that had gained popularity in probation departments not a specific prior investigation.

[3]Some of the more recent meta-analysis efforts have tried to explore the potential for isolating specific program approaches that lead to more effective outcomes (e.g., see Lipton, 1995; Sherman *et al.*, 1997).

Replication studies offer a potential for consistency and comparison that is often absent in meta-analytic studies.[4] Nonetheless, replication studies in criminal justice have generally taken an approach in which the broad parameters of treatment and method in multiple sites are similar, but specific components of study design are allowed to vary. For example, in an Institute of Policy Analysis study of restitution in four jurisdictions, the number of control and treatment conditions and their type differed in each of the jurisdictions studied (Schneider, 1986). Similarly, there was considerable variation in the nature of the programs implemented in RAND's multisite evaluation of intensive probation (Petersilia and Turner, 1993a, b). Even in the Spouse Assault Replication Program, in which the National Institute of Justice strongly encouraged the development of common program and data elements (Garner *et al.*, 1995; NIJ, 1987), there was considerable variability in protocols across sites. As Garner *et al.* (1995, p. 10) note in their review of the SARP experiments,

> These concerns [regarding analytic comparisons across sites] would be misplaced if it were possible from the published works to extract standardized or even comparable comparisons across SARP experiments. After several readings and rereadings of the published articles, books, and final reports on the SARP experiments, we were unable to do this. In fact, we could not find one comparison that could be extracted precisely for all SARP experiments.

One result of this variability between sites included in replication experiments is that overall assessments of the impacts of treatment in multisite experimental evaluations are generally qualitative. While pooled quantitative analyses of study results have been recommended by Berk *et al.* (1992b) and Garner *et al.* (1995), most investigators have chosen to weigh the results in each of the sites examined as if conclusions were being drawn from totally separate experiments. The variability across sites has generally been assumed to be too great to allow the pooling of data from the multiple sites examined. Where investigators have sought to combine multiple sites into a single analysis, they have encountered significant barriers (see Garner *et al.*, 1995) or have been forced to use a quasi-experimental rather than an experimental analytic approach (Berk *et al.*, 1992b; Sherman, 1992).

Accordingly, the approach generally taken in replication studies in criminal justice is very similar to that taken in narrative reviews summarizing multiple studies in a specific area of interest (e.g., see Farrington, 1983).

[4]It is important to note that experimental criminal justice studies that draw subjects from multiple sites or institutions are not uncommon. Petrosino (1997), for example, in a review of 150 criminal justice experiments found that fully a third drew subjects from multiple sites. However, in most of these cases, multiple sites were located in a single jurisdiction and investigators did not view site variation as a relevant characteristic in the study.

When all, or most sites, provide results of a similar strength and in a similar direction, such conclusions can be unambiguous (e.g., see Yarborough, 1979). But more often than not researchers are faced with effects of different magnitude, sometimes significant sometimes not, that lead to a degree of confusion in interpreting study results (e.g., see Garner *et al.*, 1995; Schneider, 1986).

The fact that sites in replication studies have generally been analyzed separately also has important implications for the statistical power of such experiments. Statistical power refers to the sensitivity of a research design in detecting program impacts (Lipsey, 1990; Weisburd, 1993, 1998). If a study is underpowered it is unlikely to yield a statistically significant result even if there is an effect of the treatment in the population under study. Increasing the size of the study sample is generally assumed to provide the most straightforward method for increasing the statistical power of a research design and thus avoiding the possibility that an investigation is biased toward a finding of no difference or no effect (Kraemer and Thiemann, 1987). If the researcher analyzes a multisite study separately site by site, he or she cannot take advantage of the larger number of cases that would result from pooling all of the subjects in all sites.

3. THE MULTICENTER TRIAL

In medical research, investigators have taken a different approach to multisite study. Under the rubrick of the "multicenter clinical trial" they have encouraged a method of replication study which seeks to ensure consistency of treatment and research design across multiple sites [Borok *et al.*, 1994; Fleiss, 1982; Friedman *et al.*, 1985; Hill, 1962; Stanley *et al.*, 1981; see also Greenberg *et al.* (1994) for a nonmedical example]. What distinguishes this method from traditional replication studies in criminal justice is that all sites included in a multicenter trial are considered part of one tightly controlled experimental study. It is assumed from the outset that results will be pooled, and investigators therefore begin with concern with the consistency of the study design and method that is greater than that so far applied in criminal justice replications. In this sense, a multicenter trial is a special type of replication study, in which multiple sites are not considered as separate replications but rather as part of one overall evaluation.[5]

[5]We have not mentioned another type of multisite study commonly used in evaluations of social programs. In this case, sites rather than subjects are randomly allocated to treatment and control conditions (Ellickson and Bell, 1992). For such studies, it is common to aggregate all subjects from all sites into one experimental analysis. However, since sites and not subjects are randomized, statistical controls must be brought to take into account potential correlations of site with subject variability. In this sense, such studies are not (at the subject level of analysis) true randomized experiments.

As noted earlier, the lack of a common treatment protocol has been a major impediment to the development of pooled analyses in existing multisite criminal justice studies. Often treatment varies considerably between sites, and this has been one important reason why statistical analyses have been conducted separately within each site (e.g., see Petersilia and Turner, 1993a). In multicenter clinical trials the definition of treatment is carefully laid out at the outset of a study, and each site is required to follow a common treatment protocol.[6] This means that the implementation of treatments must be carefully monitored at each site. In contrast to replication studies in criminal justice, there is an assumption at the outset that the sites must follow the same procedures in study design and implementation.

In this regard, a central problem in multicenter study is management of the large number of investigators that are likely to participate in a trial. While multicenter clinical trials may involve only a small number of centers (e.g., see the HERA Study Group, 1995), they often involve as many as 30 or more hospitals or clinics in different regions or even different countries (e.g., see Aspirin Myocardial Infarction Study Research Group, 1980). Each center often includes independent investigators that are expected to participate as equal partners in the development of the research and its dissemination. In this context, multicenter studies in medicine confront a significant organizational problem in encouraging cooperation among investigators and in ensuring consistency in study implementation. Texts describing multicenter study often focus on the management problems that such cooperation entails (e.g., see Friedman *et al.*, 1985; Pocock, 1983).

Despite the attention to consistency of program and treatment in multicenter trials, medical experimenters have recognized that treatment variation across subjects or variation in the contexts in which treatments are administered are often inevitable in multicenter studies. Variation in dosage, for example, may be an important part of clinical success. In medicine, as in criminal justice, it is often recognized that different subjects may require different levels of intervention to achieve a desired outcome (see Visher and Weisburd, 1998). This fact does not preclude multicenter evaluation but, rather, alters the research question asked by investigators from one that is concerned primarily with a specific treatment and dosage to one that defines a general process or approach that is implemented similarly at each site (see

[6]In principle, a multicenter study could be conducted in which there were common as well as different experimental treatment and control conditions in the sites studied. As long as one experimental treatment and one control condition are the same, pooled analyses can be conducted [see Garner (1995) for a suggestion of this approach in the multisite domestic violence replication studies]. Nonetheless, in practice the implementation of varied control or treatment conditions is likely to impact upon the overall comparability of the experiment at the different sites and thus the validity of the comparisons made (e.g., see Schneider, 1986).

Hill, 1962). Bradford Hill (1962, p. 11) a pioneer in multicenter clinical trials, writes in this regard:

> . . . There may well be instances in which the clinician should be allowed to adjust the dosage according to the progress and responses of the patient—to 'individualize' the treatment. That may clearly be the case in dealing with, say, rheumatoid arthritis, under treatment with cortisone. One man's meat is another man's poison. With this disease, therefore, various trials are under way in which the physician may select treatment over a wide range of dosage and from time to time . . . .

Similarly, even in multisite medical experiments which seek carefully to regulate the treatment protocol, there may be differences among hospitals or clinics in the ways in which the protocols are actually delivered, the quality of the medical personnel or facilities, or the larger physical or social environments in which the experiments take place (Fleiss, 1982). As we describe in the next section, the statistical models developed for multicenter trials allow description of such variability while, at the same time, allowing for pooled experimental analysis of study outcomes.

Recognition of variability between sites included in a multicenter experimental study is seen by medical statisticians as having important implications for understanding the effects of treatment and its use in the future. It may be the case, for example, that specific centers do better than others, a result which would suggest that there are special characteristics in those centers which interact with treatment to produce a more effective regimen. Sometimes such factors may be defined at the outset in the study design, for example, by choosing centers located in areas with very different populations that are expected to respond differently to a treatment protocol. At other times, differences may be unintended. But nonetheless, as Fleiss (1982, p. 356) notes, they may be crucial for defining the future use of a treatment.

> The reason is that the statistically deviant center or centers may have recruited into the study systematically different kinds of patients from those in the remaining centers, patients for whom the treatment that is superior on the average may actually be harmful. Their characteristics must be identified so that the treatment is not prescribed for future patients like them.

While there may be variation in the characteristics of centers included in a multicenter clinical trial, there must be consistency in implementation of the study design across sites. As Stanley *et al.* (1981, p. V) explain,

> Multi-institutional cooperative studies require greater attention to detail than studies within a single institution. For single institution studies, a single protocol document may be sufficient. In a cooperative group, however, it is necessary to standardize various aspects where little variation may be present in a single institution study. Patients must be entered in a uniform fashion, data collection and evaluation should be standardized and there must be a mechanism to insure the timely collection of essential data.

Standardization of basic components of study design in a multicenter trial allows investigators to minimize unnecessary variation between sites.

For example, while in theory different selection procedures could lead to equally valid random samples, experience in multicenter trials suggests that consistency in randomization procedures is important in minimizing variability between sites (Stanley *et al.*, 1981). In turn, the collection of consistent baseline and outcome data is essential in multicenter study (see Pocock, 1983). The absence of such consistency makes it very difficult to make comparisons across multisite studies even when the studies are analyzed independently (e.g., see Garner *et al.*, 1995). Beyond establishing common data collection and follow-up procedures, a multicenter study must have common definitions of program success and failure. In part, this is the case because outcomes are pooled by investigators. If sites define success or failure using different criteria, the meaning of the outcome evaluation becomes confused.

4. POOLING OBSERVATIONS FOR MULTICENTER STUDY: A STATISTICAL MODEL

While medical researchers have recognized the importance of characteristics that differentiate centers one from another, they have not followed the criminal justice model of analyzing multisite studies as if they comprised a series of independent replications. In good part, it is concern about statistical power that has prompted the multicenter approach (Boateng and Jones, 1994; Borok *et al.*, 1994; Friedman *et al.*, 1985; Pocock, 1983; Tygstrup, 1982). It is generally the case that a single medical facility will not be able to provide a large enough sample of patients to allow a "definitive" study (Fleiss, 1982). Accordingly, medical researchers sought to develop an approach that would allow pooling of samples across centers so that their conclusions would be based on statistically powerful research designs. Criminal justice studies often face similar limitations.

At a minimum, it is generally recommended that a statistical test have a power level greater than 0.50—indicating that the test is more likely to show a significant result than not (e.g., see Gelber and Zelen, 1985). But it is generally accepted that the most powerful experiments seek a power level of 0.80 or above (Cohen, 1988; Gelber and Zelen, 1985). Such experiments are highly likely to evidence a statistically significant finding if the expected program impact represents the real effect in the population under study. In medicine, the samples needed to ensure a statistically powerful study are sometimes small by social science standards because treatment impacts are at times expected to be strong (e.g., see Logan *et al.*, 1995). Even in these cases, multisite studies may be encouraged because very few patients with a

particular disease can be found in any one center (Hill, 1962). Nevertheless, in medicine as in other fields, most studies look to identify relatively modest effects (e.g., see Aspirin Myocardial Infarction Study Research Group, 1980; Borok *et al.*, 1994). In criminal justice the effects of treatment are also modest (Brown, 1989; Weisburd, 1993), and thus a very large number of cases is often required in order to allow a statistically powerful research design.[7]

If we assume, for example, a rearrest rate of 40% for a treatment group and 50% for a control group, using conventional significance criteria ($p < 0.05$, nondirectional research hypothesis), a sample size of about 400 cases is needed (for each treatment and control condition) to achieve a power level of 0.80.[8] Looking at prior replication studies in criminology, it is clear that researchers seldom achieve this threshold in individual sites that are included. For example, not 1 of the 14 intensive probation experiment sites examined by Petersilia and Turner (1993a) or any of the 4 restitution experiment jurisdictions studied by Schneider (1986) included enough cases to detect a program impact of this size reliably. Even though more attention was given at the outset to issues of statistical power in the domestic violence replication studies (Garner, 1987), only three of the sites included reached this threshold for analysis of official records and only one in analysis of victim interviews (Garner *et al.*, 1995). However, each of these replication studies would have reached this threshold of statistical power if the investigators had designed their studies as multicenter experiments.

In practice, the initial steps taken in developing multicenter studies are similar to those that criminal justice evaluators have used in designing replication studies. In a multicenter trial, as in most criminal justice replication studies, the investigator randomly allocates subjects "separately and independently within each center" (Fleiss, 1982, p. 354). This means that randomization is carried out as if each center were a completely separate study. But in a multicenter trial, separate randomization procedures within centers are used in order to allow statistical controls of interaction effects between treatment and treatment facility and to ensure homogeneity between the overall treatment and control comparisons when the sites are pooled into a single analysis (e.g., see Johnston *et al.*, 1994; Stangl, 1995).

The advantage of separate randomization within sites or centers is twofold. In the first case it maximizes the comparability of treatment and control conditions in the larger study. This benefit is similar to that gained in

[7]As Weisburd (1993) points out, the relationship between sample size and statistical power can be complex. Merely adding cases, without concern for the integrity of a study design, will not necessarily provide for a more statistically powerful study. Our discussion here assumes that sample size is increased without a significant impact on other aspects of the research design.

[8]For a precise estimate we would need to define the specific test being used.

block randomized experiments that match pairs or groups before randomization based on specific characteristics of the study population (see Lipsey, 1990; Sherman and Weisburd, 1995). While simple randomization of all subjects without reference to site is likely to provide an overall balance between the treatment and the control conditions, it does not guarantee that there will be complete balance on any particular trait. If the researcher knows at the outset that sites vary in significant ways, more comparable treatment and control conditions can be defined through separate randomization procedures which balance the number of cases drawn from each site, as well as the specific number of individuals placed in treatment and control conditions at each site.[9] In an approach which does not randomize separately by site, the researcher is likely to get some imbalance in the number of cases that are drawn to the larger study from any specific site.

Second, separate randomization allows the researcher to test hypotheses regarding both the pooled overall impacts of treatment and the specific impacts of treatment within separate sites, in the context of a single statistical model. The sites included in a multicenter trial can be seen in this context as building blocks that can be combined in different ways by the researcher depending on the question asked. The separate randomization procedures allow each center to be analyzed as a separate experiment as is common in multisite experimental studies. But the different sites can be combined as well into an overall experimental evaluation, in which the researcher is able to identify direct and interaction effects in a statistically powerful experimental context. If the researcher was to identify subjects from all sites and then randomly allocate subjects to treatment and control conditions (without reference to site), then any statistical analysis of the impacts of site on outcomes would be nonexperimental and subject to the same threats to internal validity due to omitted variables common in nonexperimental research (see Smith, 1990).

There are several potential models that can be used in analyzing multicenter studies and a number of choices that researchers must make regarding the nature of the factors that are examined. Ideally, a multicenter study

[9]Balance is important because it becomes much more difficult in practice to analyze study results from an unbalanced study (Lee, 1975; Iles, 1993). This is similar to the problem of randomized studies that stratify randomization by characteristics of subjects. In this case, every block, or group of subjects, should be equally divided or "balanced" in randomization between treatment and control cases. In most replication studies in criminal justice, investigators have not considered balance as an important design question because case flow into the experiment makes balance difficult to achieve. While balance between sites—that is, in the number of total cases found in each site—is not as important in multicenter studies, the more balanced the design, the more stable the study results are likely to be. Experiments that are balanced are also more likely to be robust when faced with violations of statistical assumptions of a test (Iles, 1993).

would include not only a random sample of subjects within sites examined but also a random sample of sites themselves. Without a random sample of sites, the evaluator is restricted in the types of generalizations that can be made. This choice also has implications for the type of statistical model that can be estimated. If the researcher can argue that the sites examined are a representative sample of all possible sites, then site can be assumed to be a random factor and a mixed effects model can be estimated (see Hicks, 1993; Kleinbaum *et al*., 1988). Otherwise site, must be defined as a fixed factor in the statistical models employed, and the researcher must estimate a fixed effects model. Fixed effects models allow the researcher to test whether site variation is important in the context of the sites examined but do not allow statistical statements regarding the generalizability of these effects to the larger population of study sites. They also limit testing hypotheses to the population average treatment effect, where the "population" is limited to the specific sites examined. In the mixed effects model one tests hypotheses about the average treatment effect where the population refers to the population of sites from which the sample of sites is drawn.

In practice, criminal justice evaluators are unlikely to be able to use mixed effects models in analyzing multicenter studies, both because random samples of study sites are difficult to obtain, and funding and practical constraints make it unlikely that the researcher will be able to identify enough sites to make it possible to adequately estimate such models.[10] This later point is illustrated in taking a straightforward example of a multicenter study with one treatment and one control condition, in which there are eight sites and 1600 subjects (100 subjects within each treatment group by site combination).[11] A model for this example is presented in Eq. (1). In this model treatments are crossed with sites, meaning that in each site both the treatment and the control conditions are applied. The sources of variability, degrees of freedom, and error term associated with each factor in the model are presented in Table I both for a mixed and a fixed effects model.

$$Y = \mu + S_i + T_j + ST_{ij} + O_{k(i,j)} + e_{k(i,j)} \tag{1}$$

Using this example we can see that the multicenter approach allows us to take into account both the direct sources of variability associated with treatment and site characteristics and the interaction between them. In this

[10]More often than not, researchers do not have the luxury of identifying jurisdictions in a random manner. Instead, they usually must work hard to identify sites that are willing to become involved in a randomized study. In this sense, most samples of sites are "convenience samples" rather than random samples.

[11]The HIDTA study described in the next section uses this basic research design, though an additional blocking factor is added for offender risk levels. To simplify discussion this additional factor is excluded.

Table I. Illustrating Differences Between Mixed and Fixed Effects Models (for a Study Including Eight Sites and 1600 Subjects)

(A) Source of variability	df
Site of treatment	7
Treatment type	1
Site × treatment interaction	7
Subject (nested within site × treatment interaction)	1584

(B) Degrees of freedom of error term for testing the source of variability

Source of variability	Fixed effects	Mixed effects
Site of treatment	1584	1584
Treatment type	1584	7
Site × treatment type	1584	1584

model, the researcher gains an overall estimate of the impact of treatment across the specific centers for the fixed effects model, and the larger population of centers for the mixed effects model, as represented by T_j. The estimate is based on the entire sample of cases in all of the sites studied. The average impact of a site (or center) on outcomes, averaged over treatments, is represented by the term S_i. This term captures differences, for example, in age or sociodemographic characteristics of a site that might impact on such issues as level of reoffending or quality of services irrespective of treatment. If a mixed effects model is assumed, then the researcher can make a general statement about the impacts of site characteristics on treatment. In the fixed effects model, a significant impact of S_i simply means that there is statistically significant site variation across the eight sites examined based on the sample of subjects drawn from those sites.

The term ST_{ij} represents the interaction effect between treatment and site, once the average impacts of site variation and treatment variation have been taken into account. If a mixed model is assumed, this term provides an experimental assessment of whether the impact of treatment varies across the population of sites. Using the more restricted fixed effects model, the researcher is able to identify only whether there is a treatment × site interaction in the larger population of subjects in the sites or centers examined. In practice, a finding of a statistically significant interaction would lead the researcher to speculate on specific site characteristics that might have led to more or less effective treatment impacts.

In the model proposed, subject variability, represented by the term $O_{k(i,j)}$, is nested within a site × treatment combination. While this suggests a hierarchical model, in practice $O_{k(i,j)}$ variability will be confounded with

error in the case where there is only one outcome observation on each subject. In the case where there are multiple observations on each subject, then the specific variability of these observations could be estimated as nested effects within a nested or hierarchical model (see Hicks, 1993).

Looking at the error term degrees of freedom associated with an assumption of a mixed versus fixed effects model, we can see in Table I the practical difficulty of trying to estimate a mixed model with a small number of study sites. In the case of the mixed model, to test the main effect of treatment type the appropriate error term degrees of freedom is 7 (corresponding to the source of variability associated with the site and treatment interaction). If there is a meaningful interaction between site and treatment, then it will become difficult to gain a statistically significant F statistic. However, if we assume a fixed effects model, the degrees of freedom for the error term for testing treatment effects is 1584 (corresponding to the source of variability associated with the subject, nested within site and treatment interaction). While there is no specific requirement regarding the number of sites that would be needed to overcome such estimation difficulties, Greenberg *et al.* (1993) suggest that multisite studies should include no fewer than 20 and sometimes well over 100 sites.

As this example illustrates, the multicenter model will allow for a statistically powerful research design for analyzing overall treatment impacts at the same time that it provides a method for taking into account the importance of intersite variation on experimental outcomes. While this design has specific advantages compared to presently used methods, it demands that researchers address a number of design and analysis issues that are often ignored in multisite criminal justice studies.

Below we examine such issues in the context of the Washington/Baltimore HIDTA (High-Intensity Drug Trafficking Areas) Project and Multi-Center Evaluation. Our discussion centers on two main questions:

(1) How can program elements be defined in a criminal justice study in order to facilitate a multicenter trial?
(2) What features of study implementation should be addressed in order to allow for valid pooling of experimental sites?

5. THE WASHINGTON/BALTIMORE HIDTA PROJECT AND MULTICENTER EVALUATION

As part of its general mandate to reduce drug consumption among substance abusing offenders and to provide treatment services, the Office of National Drug Control Policy initiated a demonstration project focusing on the development of a "seamless" system approach to drug-involved

offenders that would combine criminal justice supervision and drug treatment in multiple jurisdictions in the Washington, DC–Baltimore corridor (see Taxman and Lockwood, 1996). A seamless system is a service delivery system which links criminal justice and treatment agencies together with umbrella policies and practices and thus reduces the risk that offenders will fall through the cracks that are often created when they shift from the responsibility of one agency to another (see Section 6 for a fuller description of the seamless system approach used in HIDTA). Twelve jurisdictions received HIDTA funding for a testing, treatment, and sanctions protocol. Of these, eight were identified for possible inclusion in a randomized evaluation based on their implementation of HIDTA initiatives and the inclusion of more than minimal HIDTA case load.[12] The goals of the HIDTA project are to increase client participation in treatment, to reduce substance abuse, to reduce criminal behavior, and to improve social adjustment.

Evaluators recognized at the outset that the HIDTA project provided a unique opportunity for developing a large randomized evaluation of the seamless criminal justice system model. Researchers had worked with practitioners in the development of the HIDTA approach, a process that created an atmosphere of trust between researchers and practitioners that was likely to facilitate cooperation in implementing a randomized study. Such cooperation has been found to be a major factor in implementation of prior randomized experiments in criminal justice (e.g., see Petersilia, 1989). Moreover, the fact that there were many more offenders in participating jurisdictions who were eligible for HIDTA treatment than could be included in the HIDTA program lessened ethical concerns of practitioners and, thus, made it easier to gain the consent of participating agencies for use of randomization to select clients for drug treatment.

While the HIDTA program provided a conducive environment for the development of a randomized study, it was apparent from the outset that it would be difficult to develop a statistically powerful evaluation of HIDTA in any specific site. A review of prior treatment studies suggested that

[12]The participating jurisdictions include Charles County; Prince William County; Fairfax County; Baltimore City; Montgomery County; Baltimore County; Washington, DC; and Alexandria City. Four sites had begun randomization by November of 1999. Additional sites are added based on their progress in implementing the seamless system components using a procedure based on the integrity of the protocol and evidence of the ability to successfully implement randomization. A three-step process is used to determine implementation of the experiment—continuum of care is in place, drug testing occurs, and the probation staff use a sanction protocol. The site must also have a risk tool that differentiates among high- and moderate/low-risk offenders. Most of the jurisdictions did not have a system for screening offenders for treatment prior to the HIDTA program—instead services were offered on a first come/first served basis. Creation of such a system was essential for developing randomization protocols in the study.

impacts of treatment were likely to be modest (e.g., see Andrews *et al.*, 1990; Lipton, 1995). While evaluators hoped that the seamless approach would strengthen treatment impacts, they did not want to design a study that could not identify a small program effect. A study of case flow in each of the eight experimental jurisdictions showed that only one individual site could be expected to have more than 50 treatment cases in any given month.

Moreover, the investigators recognized that the randomized evaluation would be difficult to implement if the randomization process went on for too long [see Dennis (1990) for a more general discussion of the problems associated with changes in the "environmental context" of an experiment over time]. Practitioners in the sites could not be expected to allocate treatment on the basis of randomization indefinitely. Funding limitations also meant that the initial sample would have to be developed as quickly as possible, in order to allow sufficient follow-up of study subjects. Based on these concerns, it was decided that a maximum of 6 months could be used as a sample selection period. Using this estimate, only one of the participating sites would gain enough cases for a statistically powerful evaluation of small program impacts.

As in other multicenter studies, the need for a larger number of cases prompted the researchers to consider the multicenter approach. If all sites could be pooled into a single study, statistical power concerns could be overcome. In turn, the multicenter approach would allow investigators to develop an overall measure of the importance of the interaction between study site and treatment. HIDTA researchers, however, as other criminal justice evaluators, faced an important hurdle at the outset in defining a multicenter approach. Could a common program design be developed in a series of different sites that would allow researchers to define a common treatment protocol across all centers?

6. DEFINING A COMMON PROTOCOL FOR EVALUATION

The common protocol of the Washington/Baltimore HIDTA criminal justice and treatment initiative is found in the seamless criminal justice system approach. The seamless service delivery system links criminal justice and treatment agencies together with umbrella policies and practices (Taxman and Lockwood, 1996; see also, Moore, 1992). In a seamless system these umbrella policies and practices guide the actions of the agencies. Under the seamless approach, treatment and criminal justice agencies work in tandem to improve outcomes of HIDTA offenders. That is, instead of each agency acting on its own, as is the common practice in drug treatment and supervision (Duffee and Carlson, 1996; Swartz *et al.*, 1996), the agencies coordinate efforts and combine resources to provide what is defined as an

appropriate level of services. Unlike individual case management, where services are brokered and obtained on an as-available basis, the focus is on ensuring that the treatment and criminal justice agencies have agreed on a service delivery system that guides the amount and type of services delivered in each agency.

In order to provide a common treatment protocol in the HIDTA project, standard levels of care for offenders were defined as a minimum of two consecutive treatment experiences (for a total of 12 months of treatment), appropriate criminal justice supervision, a minimum of three times a month drug testing, and the implementation of graduated sanctions or swift and certain responses to offender noncompliance. The sanctions become a vehicle to adjust treatment and supervision services to offender progress levels. Overall, the protocol is designed on "best practices" with a focus on the goal of increasing the length of stay in treatment as a means to improve offender outcomes (Lipton, 1995).

Program evaluators recognized at the outset the importance of linking the HIDTA approach to an established protocol of treatment that could be implemented in similar ways across the sites included in the evaluation [see Boruch (1997), Dennis (1990), Petersilia (1989), and Weisburd (1993) for a discussion of the difficulties of ensuring treatment integrity in experimental research]. In this context, three phases of HIDTA treatment were defined. In the first phase offenders experience residential treatment including detoxification and group and individual therapy for between 30 and 45 days. In the second stage (lasting 6 months), HIDTA subjects are provided with intensive outpatient care in which group and individual therapy continue, and the offenders are routinely tested for drug involvement. Finally, for a 6-month "aftercare" phase, offenders continue drug testing, treatment, and supervision at a reduced level. Graduated sanctions are also imposed (when required) during all phases of treatment in order to enforce compliance of treatment conditions.

Each HIDTA site has followed the basic protocol of the seamless system. However, as Table II indicates, in practice individual differences do exist between sites as well as across offenders. For example, while the treatment consists of three levels of care for at least 13 months in all sites, two sites carry out part of the treatment in jail and then continue treatment in the community.[13] While all sites have a similar "dosage" of treatment (in

[13]While the identification of subjects in different criminal justice settings raises a question as to whether the types of offenders examined across the sites are similar, a preliminary analysis of 1700 cases across HIDTA sites suggests that the population pools are similar in terms of criminal justice history, substance abuse history, and prior treatment experience. Moreover, the HIDTA protocol targets "hard-core offenders." Similar definitions are used for identifying such offenders across sites.

Table II. Core Features of the HIDTA Seamless System Approach

Assessment/determination of hard-core offenders
• Addiction Severity Index (ASI)
• Probation needs/risk instrument
Level I: Placed in HIDTA residential treatment (approximately 30 to 45 days)[a]
All sites provide: 20–30 h per week of cognitive behavior therapy (e.g., detoxification, group therapy, individual therapy, and transition planning)
Setting of treatment
Jail (Charles County, Prince William County, Fairfax County, Baltimore City)
Special facility (Montgomery County, Baltimore County)
Residential/contractual services (Washington, DC, Alexandria City)
Level II: Intensive outpatient care (approximately 6 months)
All sites provide: 20–30 h per week of cognitive behavior therapy (e.g., group therapy, individual therapy, and aftercare planning)
Setting of treatment
Probation office (Prince William County, Fairfax County, Washington, DC, Alexandria City, Baltimore City)
Treatment facility (Montgomery County, Baltimore County)
Supervision: Intensively supervised at least twice a week
Drug testing
2 times per week (Washington, DC, Montgomery County, Baltimore City)
3 times per month (Alexandria City, Charles County, Baltimore County, Prince William County, Fairfax County)
Graduated sanctions for noncompliance
All sites include a progressive set of sanctions (that follow a pattern similar to the following):
1. Three-way conference among supervision agents, treatment providers, and offenders
2. Increased drug testing, reporting, curfews, and therapy
3. Short-term jail stay, residential treatment, electronic monitoring.
Level III: Outpatient therapy/aftercare (approximately 6 months)
All sites provide: Once-a-week support group and outpatient treatment
Drug testing continued once per month
Supervision on regular probation

[a]Sites allow immediate acceptance into Level II if offenders are not considered public safety risks or do not have a jail sentence.

terms of the number of total hours of service), treatments are administered in both outpatient clinics and probation offices. At the offender level, the HIDTA approach was defined as including three distinct phases. However, individuals in any of the sites may be placed directly in the second stage of treatment if they are defined as not in need of residential or structured intensive treatment. Moreover, aspects of treatment such as graduated sanctions vary by offender depending on their behavior during the experiment,

and only those who violate aspects of the program will actually be subject to graduated sanctions.[14]

Accordingly, while the HIDTA experiment has been designed to maximize the consistency of treatments across sites and offenders, the program varies depending on the needs of offenders and the context of the sites in which treatment is delivered. Such variability will be captured in part through the statistical model described above. However, the potential for significant variability across subjects and sites suggests that careful monitoring of treatment be included. Following this the HIDTA project is collecting detailed data monthly on drug testing, supervision, social services, and treatment services. These data are being used to establish when a site is implementing the seamless system protocol with enough consistency to begin the experimental period of operation, as well as to ensure that the sites are maintaining the minimum requirements for the study protocol once the experimental period has begun (see also Dennis, 1990; Sherman and Weisburd, 1995).

Data on treatment processes are also providing important information regarding what treatments are actually delivered at the eight HIDTA sites. In HIDTA, the question being asked is not whether a specific dosage or treatment is effective in improving offender outcomes but whether the HIDTA approach overall improves those outcomes. In prior experiments the "black box" of treatment has often hindered the ability of evaluators to define what it is that has produced or not produced a significant program impact (Martinson, 1974). In this case, detailed data collected about interventions will allow evaluators to define clearly the nature of the treatment and control interventions and to identify differences among sites that could account for a significant treatment $\times$ site interaction.

Our discussion so far has focused on the definition of HIDTA treatment. In multicenter experimental study there must also be a common protocol for the control condition. Recognizing the importance of clearly defining the control or comparison group, the study evaluators worked with each site in identifying what types of interventions may be provided to control group subjects. The control group in this context was defined as receiving "traditional criminal justice supervision services." The decision not to

[14]Graduated sanctions are tools used by criminal justice and treatment agencies to address noncompliance. These involve structured responses that are delivered shortly after an infarction. A typical response might include increased supervision, drug testing, or treatment sessions. Each jurisdiction has developed sanctions that fit within the sociolegal climate in that jurisdiction. The delivery effort may vary because each court may have different preferences about the degree to which supervision agents can administer sanctions. In some systems, special court dockets are provided, while others involve behavioral contracts that are approved by the Judiciary.

rely upon a "placebo" or no-intervention control conditions derives from both substantive and public safety concerns. Substantively, our interest is in whether the HIDTA seamless system is an improvement over traditional criminal justice supervision (which generally involves little treatment or drug education), not whether HIDTA works better than no intervention at all. Regarding public safety, all sites are legally required to provide some type of criminal justice supervision or intervention for drug active offenders since these offenders are under correctional control. In other experimental studies, treatments are also commonly compared to existing protocols (e.g., see Weisburd and Green, 1995; Zhang, 1996).

While investigators recognized that the control condition, as the experimental condition, must be allowed to vary somewhat by site and offender depending on the nature of the site and the behaviors of the offender, a set of common criteria was also set for "treatment" of control subjects. Typically this includes face-to-face and collateral contacts with probation officers and some limited drug testing. Control subjects are not provided with therapeutic interventions nor a continuum of care. This means in practice that offenders in the control group are not placed in group or individual therapy, detoxification, residential or inpatient treatment, jail-based treatment, halfway houses, or day reporting programs.[15]

7. ENSURING CONSISTENCY IN IMPLEMENTATION OF THE STUDY DESIGN ACROSS SITES

Unlike many other multicenter studies, one team of investigators is responsible for implementing the HIDTA experiments in all sites. While the fact that there is only one research team simplifies management of the investigation, there are still monthly meetings in which program managers from all sites are brought together to discuss elements of treatment and project design. The importance of such meetings for ensuring the consistency of treatment implementation has been illustrated a number of times in the development of the experiment. Early on in the design of the study practitioners had voiced concerns regarding how the project would be represented to potential participants. The monthly project meetings provided a venue for explaining to staff at all sites how the process of informed consent worked and enabled practitioners to discuss collectively potential problems that might arise. This ensured that study protocols were implemented in the

[15]To implement the experiment, the sites agreed to refrain from referring control group members to treatment programs. However, once the control group member becomes a research "failure" (e.g., the offender has three or more technical violations), the agency can reassign the offender to treatment services, if appropriate.

same way in each site. Another example of the importance of monthly meetings develops from the realities of staff changes in probation and drug treatment agencies. New staff are brought to monthly meetings to brief them on the experimental protocols as they are implemented in the various sites. This provides an opportunity for new staff to voice concerns and for veterans in the experiment to reinforce the nature of treatment protocols and the fact that they can be implemented successfully.

Because HIDTA began as a multicenter study, the development of common data collection instruments and procedures was required at the outset. A series of instruments is administered by research staff at the time of random assignment. The protocol requires that research staff first request consent to participate in the study, then administer the research instruments, and, finally, make the randomization assignment.[16] Program staff are informed of the random assignment before they complete their own placement interview with the offender. The interview is approximately 2 hours in duration with the following instruments being administered: the Addiction Severity Index (ASI), the Criminal Justice Experience, HIV/AIDS Risk Behaviors, Treatment Experience Survey, and Criminal/Substance Abuse Calendar. Each instrument is administered at onset as well as at 12-month intervals at every site.

Beyond establishing common data collection and follow-up procedures, a multicenter study must have common definitions of program success and failure. If different sites define failure using different criteria, the meaning of the outcome evaluation becomes confused. But common definitions of success also relate to the progression of offenders through treatment. To ensure that HIDTA and control group subjects have an "equal risk for failure," participating sites agreed to impose drug testing at the same rate for both groups. Further, program completion is defined by each jurisdiction similarly during the course of the experiment.

An offender is classified as a "success" if he or she completes the treatment and supervision regime with no more than three administrative or graduated sanctions during the supervision period. The sites agreed that offenders will be "technically violated" in both treatment and control groups if they have more than three noncompliance incidents and no change in behavior as a result of the graduated sanctions. This is particularly important because technical violation may lead to reincarceration and thus

[16]The experimental design also requires blocking by risk level. Risk information is provided by the criminal justice staff to the researchers. This information is then used to place offenders in block randomized groups. The assignment protocol takes into consideration the risk level to ensure that there is a balanced design including an equal number of offenders assigned to treatment and control groups by risk level. This will allow testing of hypotheses about the effectiveness of the intervention for offenders at different risk levels.

would end treatment of offenders and preclude risk of reoffending. However carefully protocols are defined, the involvement of multiple sites in a study is likely to lead to unintended variability. For example, at one site probation officers accelerated the "failure" of subjects in order to expedite their placement in drug treatment programs (even though in reality there were no available places for treatment of these subjects). At this site, investigators stopped the experiment and began again only when practitioners had agreed to follow the standard protocol.

8. CONCLUSIONS

Multicenter clinical trials allow evaluators to conduct statistically powerful experimental studies in which solid overall conclusions about progam impacts can be made. Multicenter studies also enable researchers to examine the impacts of different subject populations or treatment settings on the outcomes of an experimental study. In this paper we have argued that the model of a multicenter trial can be applied successfully to multisite criminal justice experiments. Nonetheless, criminal justice studies are likely to face significant problems in defining regimented treatments, in ensuring treatment integrity and consistency, and in developing and implementing common protocols and methods across sites. We have discussed some of these issues as they relate to the HIDTA evaluation, arguing that such difficulties can be overcome. However, the potential of multicenter criminal justice study as well as the special problems it presents will become better understood only after such methods are more widely applied in criminal justice.

ACKNOWLEDGMENTS

Support for this research has been provided under cooperative agreement I9PWBP528 from the Office of National Drug Control Policy and grant award 96CEVX0017 from the National Institute of Justice. Points of view in this paper are those of the authors and do not necessarily represent the official positions of these federal agencies. We would like to thank Joe Naus for his helpful advice on statistical questions; Anthony Petrosino, Joel Garner, and Iaain Chalmers for their thoughts regarding experimental study; and Bruce Kubu, Christine DeFastano, Dorothy Lockwood, Rebecca Silverman, and Meredith Thanne for their assistance in implementing the project. Helpful suggestions to improve the paper were also provided by Michael Maltz and anonymous reviewers of *JQC*.

REFERENCES

Abrami, P. C., Cohen, P. A., and d'Apollina, S. (1988). Implementation problems in meta-analysis. *Rev. Educ. Res.* 58(2): 151–179.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., and Cullen, F. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology* 28(3): 369–404.

Aspirin Myocardial Infarction Study Research Group (1980). A randomized controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 243: 661–669.

Barton, W., and Butts, J. (1990). Viable options: Intensive supervision programs for juvenile delinquents. *Crime Delinq.* 36: 238–256.

Berk, R., Campbell, A., Klap, R., and Western, B. (1992a). A Bayesian analysis of the Colorado Springs Spouse Abuse Experiment. *J. Crim. Law Criminol.* 83: 174–201.

Berk, R., Campbell, A., Klap, R., and Western, B. (1992b). The deterrent effect of arrest in incidents of domestic violence: A Bayesian analysis of four field experiments. *Am. Sociol. Rev.* 57(5): 698–708.

Binder, A., and Meeker, J. W. (1988). Experiments as reforms. *J. Crim. Just.* 16: 347–358.

Boateng, J., and Jones, D. (1994). An evaluation of six new intrauterine devices. *Adv. Contracept.* 10(1): 57–70.

Borok, G., Reuben, D., Zendle, L., Ershoff, D., Wolde-Tsadik, G., Rubenstein, L., Ambrosini, V., Fishman, L., and Beck, J. (1994). Rationale and design of a multi-center randomized trial of comprehensive geriatric assessment consultation for hospitalized patients in an HMO. *J. Am. Geriatr. Soc.* 42: 536–544.

Boruch, R. F. (1997). *Randomized Experiments for Planning and Evaluation*, Sage, Newbury Park, CA.

Brown, S. E. (1989). Statistical power and criminal justice research. *J. Crim. Just.* 17: 115–122.

Campbell, D., and Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Chicago, IL.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Cook, T. D., Cooper, H., Cordray, D., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., and Mosteller, F. (1992). *Meta-Analysis for Explanation: A Casebook,* Russell Sage Foundation, New York.

Dennis, M. L. (1990). Assessing the validity of randomized field experiments: An example from drug abuse treatment research. *Eval. Rev.* 14(4): 3457–373.

Duffee, D. E., and Carlson, B. E. (1996). Competing value premises for the provision of drug treatment to probationers. *Crime Delinq.* 42(4): 574–592.

Ellickson, P. L., and Bell, R. M. (1992). Challenges to social experiments: A drug prevention example. *J. Res. Crime Delinq.* 29: 79–101.

Etheridge, R. M., Hubbard, R. L., Anderson, J., Craddock, S. G., and P. M. Flynn (1997). Treatment structure and program services in the Drug Abuse Treatment Outcome Study (DATOS). *Psychol. Addict. Behav.* 11(4): 244–261.

Eysenck, H. J. (1978). An exercise in mega-silliness. *Am. Psychol.* 33: 517.

Farrington, D. P. (1983). Randomized experiments on crime and justice. In Tonry, M., and Morris, N. (eds.), *Crime and Justice: An Annual Review of Research, Vol. 4*, University of Chicago Press, Chicago, pp. 257–308.

Farrington, D. P., Ohlin, L. E., and Wilson, J. Q. (1986). *Understanding and Controlling Crime*, Springer Verlag, New York.

Fleiss, J. (1982). Multicentre clinical trials: Bradford Hill's contributions and some subsequent developments. *Stat. Med.* 1: 353–359.

Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*, John Wiley and Sons, New York.

Friedman, L., Furberg, C., and DeMets, D. (1985). *Fundamentals of Clinical Trials*, PSG, Littleton, MA.

Garner, J. (1987). Second phase funding Milwaukee replication, Unpublished memorandum, National Institute of Justice, Washington, DC.

Garner, J. (1990). Two, three . . . many experiments. The use and meaning of replication in social science research. Paper presented at the Annual Meeting of the American Society of Criminology, Baltimore, Nov.

Garner, J., Fagan, J., and Maxwell, C. (1995). Published findings from the Spouse Assault Replication Program: A critical review. *J. Quant. Criminol.* 11: 3–28.

Gelber, R., and Zelen, M. (1985). Planning and reporting clinical trials. In Calabrese, P., Schein, P., and Rosenberg, S. (eds.), *Basic Principles and Clinical Management of Cancer*, Macmillan, New York.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educ. Res.* 5: 3–8.

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-Analysis in Social Research*, Sage, Beverly Hills, CA.

Greenberg, B. G. (1959). Conduct of cooperative field and clinical trials. *Am. Stat.* 13: 13–28.

Greenberg, D., Meyer, R. H., and Wiseman, M. (1993). Prying the lid from the black box: Plotting evaluation strategy for welfare employment and training programs, Discussion paper. Institute for Research on Poverty, University of Wisconsin–Madison, Madison.

Greenberg, D., Meyer, R. H., and Wiseman, M. (1994). Multisite employment and training program evaluations: A tale of three studies. *Industr. Labor Relat. Rev.* 47(4): 679–691.

HERA Study Group (1995). A randomized trial of hydroxychloroquine in early rheumatoid arthritis: The HERA study. *Am. J. Med.* 98: 156–168.

Hicks, C. R. (1993). *Fundamental Concepts in the Design of Experiments*, 4th ed., Saunders College, San Francisco.

Hill, B. (1962). *Principles of Medical Statistics*, Oxford University Press, New York.

Hutchison, I. W., and Hirschel, J. D. (1994). Limitations in the pro-arrest responses to spouse abuse. *J. Contemp. Crim. Just.* 10: 147–163.

Iles, T. C. (1993). Crossed and hierarchical analysis of variance. In Fry, J. C. (ed.), *Biological Data Analysis: A Practical Approach,* Oxford University Press, New York.

Johnston, P., Coniff, R., Hoogwerf, B., Santiago, J., Pi-Sunyer, F., and Krol, A. (1994). Effects of the carbohydrase inhibitor miglitol in sulfonylurea-treated NIDDM patients. *Diabetes Care* 17: 20–29.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, PWS-Kent, Boston.

Kraemer, H., and Thiemann, S. (1987). *How Many Subjects: Statistical Power Analysis in Research*, Sage, Newbury Park, CA.

Lee, W. (1975). *Experimental Design and Analysis*, W. H. Freeman, San Francisco.

Light, R. J., and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing*, Harvard University Press, Cambridge, MA.

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research,* Sage, Newbury Park, CA.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In Cook T. D. *et al.* (eds.), *Meta Analysis for Explanation: A Casebook,* Russell Sage Foundation, New York, pp. 83–127.

Lipton, D. (1995). *The Effectiveness of Treatment for Drug Abusers Under Criminal Justice Supervision*, NIJ Research Report, National Institute of Justice, Washington, DC.

Logan, C. H., and Gaes, G. G. (1993). Meta-analysis and the rehabilitation of punishment. *Just. Q.* 10(2): 245–264.

Logan, C. H. *et al.* (1995). Eradication of Helicobacter pylori and prevention of recurrence of duodenal ulcer: A randomized, double-blind, multi-centre trial of omeprazole with or without clarithromycin. *Aliment Pharmacol Ther.* 9(4): 417–423.

MacKenzie, D., and Souryal, C. (1994). *Multisite Evaluation of Shock Incarceration: Evaluation Report*, Evaluation Report to the National Institute of Justice, National Institute of Justice, Washington, DC.

Martinson, R. (1974). What works? Questions and answers about prison reform. *Public Interest* 35: 22–54.

McShane, M., Williams, F., and Wagoner, C. (1992). Prison impact studies: Some comments on methodological rigor. *Crime Delinq.* 38: 105–120.

Moore, S. T. (1992). Case management and the integration of services: How service delivery systems shape case management. *Social Work* 37: 418–423.

National Institute of Justice (NIJ) (1985). *Replicating an Experiment in Specific Deterrence: Alternative Police Responses to Spouse Assault,* National Institute of Justice, Washington, DC.

National Institute of Justice (NIJ) (1987). Common data elements, Unpublished memorandum, Spouse Assault Replication Program, National Institute of Justice, Washington, DC.

Petersilia, J. (1989). Implementing randomized experiments: Lessons from BJA's intensive supervision project. *Eval. Rev.* 13: 435–458.

Petersilia, J., and Turner, S. (1990). *Intensive Supervision for High Risk Probationers: Findings from Three California Experiments*, RAND, Santa Monica, CA.

Petersilia, J., and Turner, S. (1993a). Intensive probation supervision. *Crime Just.* 17: 281–335.

Petersilia, J., and Turner, S. (1993b). *Evaluating Intensive Supervision Probation/Parole: Results of a Nationwide Experiment,* National Institute of Justice Research in Brief, Washington, DC.

Petrosino, A. J. (1995). Specifying inclusion criteria for a meta-analysis: Lessons and illustrations from a quantitative synthesis of crime reduction experiments. *Eval. Rev.* 19(3): 274–293.

Petrosino, A. J. (1997). *"What Works?" Revisited Again: A Meta-Analysis of Randomized Experiments in Individual-Level Interventions,* Unpublished dissertation, School of Criminal Justice, Rutgers University, New Brunswick, NJ.

Pocock, S. (1983). *Clinical Trials: A Practical Approach,* John Wiley and Sons, New York.

Sackett, D. L., and Rosenberg, W. C. (1995). On the need for evidence based medicine. *Health and Economics* 4: 249–254.

Schneider, A. (1986). *Restitution and recidivism rates of juvenile offenders*: *Results from four experimental studies*. *Criminology* 24(3): 533–552.

Sechrest, L., and Rosenblatt, A. (1987). Research methods. In Quay, P. H. C. (ed.), *Handbook of Juvenile Delinquency*, John Wiley and Sons, New York, pp. 417–450.

Sherman, L. W. (with Schmidt, J., and Rogan, D.) (1992). *Policing Domestic Violence: Experiments and Dilemmas*, Free Press, New York.

Sherman, L. W. (1998) *Evidence Based Policing*, Series on Ideas in American Policing, Police Foundation, Washington, DC.

Sherman, L. W., and Berk, R. (1984). *The Minneapolis Domestic Violence Experiment*, Police Foundation Reports, No. 1, Washington, DC.

Sherman, L. W., and Cohen, E. G. (1989). The impact of research on legal policy: The Minneapolis Domestic Violence Experiment. *Law Soc. Rev.* 23: 117–144.

Sherman, L. W., and Weisburd, D. (1995). General deterrent effects of police patrol in crime 'hot spots': A randomized controlled trial. *Just. Q.* 12(4): 625–648.

Sherman, L. W., Smith, D., Schmidt, J., and Rogan, D. (1992). Crime, punishment and stake in conformity: Legal and informal control of domestic violence. *Am. Sociol. Rev.* 57: 680–690.

Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising: A Report to the United States Congress*, National Institute of Justice, Washington, DC.

Smith, H. L. (1990). Specification problems in experimental and nonexperimental social research. In Clogg, C. C. (ed.), *Sociological Methodology*, American Sociological Association, Washington, DC.

Stangl, D. (1995). Prediction and decision making using Bayesian hierarchical models. *Stat. Med.* 14: 2173–2190.

Stanley, K., Stjernsward, M., and Isley, M. (1981). *The Conduct of a Cooperative Clinical Trial*, Springer-Verlag, New York.

Swartz, J. A., Lurigio, A. J., and Slomka, S. A. (1996). The impact of IMPACT: An assessment of the effectiveness of a jail-based treatment program. *Crime Delinq.* 42(4): 553–573.

Taxman, F. (1998). *Reducing Recidivism Through a Seamless System of Care: Components of Effective Treatment, Supervision, and Transition Services in the Community,* Office of National Drug Control Policy, Washington, DC.

Taxman, F., and Lockwood, D. (1996). *Systemic Case Management: The Washington–Baltimore High Intensity Drug Trafficking Areas Treatment and Criminal Justice Supervision Project,* Unpublished paper, University of Maryland, College Park.

Tilly, N. (1994). *After Kirkhold—Theory, Method and Results of Replication Evaluations*, Police Research Group, Crime Prevention Unit Series Paper No. 47, Home Office Police Department, London.

Toborg, M. A., Sorin, M. A., and Pyne, D. (1979). *The Outcomes of Pretrial Release: Preliminary Findings of the Phase II National Evaluation*, Lazar Institute, Washington, DC.

Tygstrup, N. (1982). Achieving adequate sample size: The multicenter trial. In Tygstrup, N., Lachin, J. M., and Juhl, E. (eds.), *The Randomized Clinical Trial and Therapeutic Decisions*, Marcel Dekker, New York and Basel.

U.S. Attorney General's Task Force on Family Violence (1984). *Final Report*, Government Printing Office, Washington, DC.

Visher, C. A., and Weisburd, D. (1998). Identifying what works: Recent trends in crime prevention strategies. *Crime Law Soc. Change* 28(3–4): 223–242.

Weisburd, D. (1993). Design sensitivity in criminal justice experiments. *Crime Just.* 17: 337–339.

Weisburd, D. (1998). *Statistics in Criminal Justice*, Wadsworth, Belmont, CA.

Weisburd, D., and Greene, L. (1995). Policing drug hot spots: The Jersey City Drug Market Analysis Experiment. *Just. Q.* 12: 711–735.

Whitehead, J. T., and Lab, S. P. (1989). Meta-analysis of juvenile correctional treatment. *J. Res. Crime Delinq.* 26(3): 276–295.

Yarborough, J. (1979). *Evaluation of JOLT as a Deterrence Program*, Michigan Department of Corrections, Lansing.

Zhang, S. X. (1996). The efficiency of working under one roof: An evaluation of Los Angeles County juvenile justice centers. *Crime Delinq.* 42(2): 224–257.